

Unbiased estimators for spatial distribution functions of classical fluids

Artur B. Adib^{1,2,*} and Christopher Jarzynski^{1,†}

¹ *Theoretical Division, T-13, MS B213,
Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

² *Department of Physics, Box 1843, Brown University,
Providence, Rhode Island 02912, USA*

(Dated: February 2, 2008)

Abstract

We use a statistical-mechanical identity closely related to the familiar virial theorem, to derive unbiased estimators for spatial distribution functions of classical fluids. In particular, we obtain estimators for both the fluid density $\rho(r)$ in the vicinity of a fixed solute, and for the pair correlation $g(r)$ of a homogeneous classical fluid. We illustrate the utility of our estimators with numerical examples, which reveal advantages over traditional histogram-based methods of computing such distributions.

LAUR-04-6270

*Electronic address: artur@brown.edu

†Electronic address: chrisj@lanl.gov

I. INTRODUCTION

In molecular simulations of fluids, one is often interested in estimating a *radial distribution function* $\rho(r)$, that is, the average density of particles as a function of distance r away from either a tagged reference particle (the pair correlation function of uniform fluids[1]), or a fixed point in space (useful in the context of nonuniform fluids[2]). Almost invariably, such estimates are constructed using histograms, with $\rho(r)$ obtained by counting particle visits to a spherical shell of radius r and thickness Δr centered about the point of interest [1, 3]. While simple to implement, this approach inevitably introduces systematic error, or *bias*, as the intended estimator of $\rho(r)$ is in reality an estimator of the particle density averaged over the finite thickness Δr of the shell. This bias can of course be made small by reducing Δr , but only at the price of increased statistical error: the smaller the bin size Δr , the fewer the particles found in a typical shell.

In this paper we show how to construct *unbiased* estimators of radial distribution functions. Upon ensemble averaging these estimators coincide with the desired observables, thus eliminating the systematic error that arises in the histogram methods, and altogether avoiding the problem of finite bin size.

Our estimators express a local, surface quantity in terms of a volume average; for a given r , contributions to the estimate of $\rho(r)$ come from particles throughout the bulk of the fluid, and not only from those in a thin spherical shell at r . This approach is similar to the widely used virial method for computing the pressure p of a fluid[4]: while p can be defined operationally as the force per unit area exerted by the fluid on the walls of its container (a surface quantity), the virial theorem re-expresses p as a sum of the ideal gas pressure and an exact correction given in terms of particle-particle forces throughout the bulk of the fluid.

In what follows we first introduce the identity that underlies our approach (Eq.3), then we apply this identity to obtain estimators of commonly studied radial distribution functions (Eqs.9, 16, 19, 20, 24). We present numerical evidence illustrating the practical utility of our approach in the context of a Lennard-Jones fluid, and we conclude with a discussion of potential extensions applicable to more elaborate situations.

II. THEORY

Consider a fluid of N identical, classical particles confined within a cubic simulation box of dimensions $L \times L \times L$ and volume $V = L^3$, with periodic boundary conditions.¹ Let $\Phi(\mathbf{r}_1, \dots, \mathbf{r}_N)$ be the potential energy expressed in terms of the locations of the particles. This potential includes both particle-particle interactions and nonuniform terms such as the interaction of each particle with a fixed solute. The density $\rho(\mathbf{r})$ in the canonical ensemble is given by:[3]

$$\rho(\mathbf{r}) = \frac{1}{Z_r} \sum_{i=1}^N \int_V d\mathbf{r}^{(-i)} e^{-\beta\Phi(\mathbf{r}_1, \dots, \mathbf{r}_N)} \Big|_{\mathbf{r}_i=\mathbf{r}}, \quad (1)$$

where $\int_V d\mathbf{r}^{(-i)}$ denotes an integration over the system volume V and all coordinates except \mathbf{r}_i , and $Z_r \equiv \int d\mathbf{r}_1 \cdots d\mathbf{r}_N e^{-\beta\Phi}$ the configurational part of the partition function. Let us now choose an arbitrary three-dimensional vector field $\mathbf{u}(\mathbf{r})$, and apply the divergence (Gauss's) theorem to the product $\rho\mathbf{u}$:

$$\oint_{\partial\Omega} d\mathbf{a} \cdot \mathbf{u} \rho = \int_{\Omega} d\mathbf{r} \rho \nabla \cdot \mathbf{u} + \int_{\Omega} d\mathbf{r} \mathbf{u} \cdot \nabla \rho. \quad (2)$$

The two terms on the right are integrals over an arbitrary volume Ω contained entirely within the simulation box; the term on the left is an integral over the surface $\partial\Omega$ of this volume. Eqs.1 and 2 combine to give

$$\oint_{\partial\Omega} d\mathbf{a} \cdot \mathbf{u} \rho = \int_{\Omega} d\mathbf{r} \rho \nabla \cdot \mathbf{u} + \beta \left\langle \sum_{i \in \Omega} \mathbf{u}_i \cdot \mathbf{F}_i \right\rangle, \quad (3)$$

where the sum is over all particles inside the domain Ω , the notation \mathbf{u}_i is shorthand for $\mathbf{u}(\mathbf{r}_i)$, and $\mathbf{F}_i = -\nabla_i \Phi$ is the net instantaneous force on particle i . The angle brackets denote the canonical average, $\langle \cdots \rangle = Z_r^{-1} \int d\mathbf{r}_1 \cdots d\mathbf{r}_N e^{-\beta\Phi} \cdots$.

In the examples discussed below, the force \mathbf{F}_i on a given fluid particle includes contributions from interactions with the other $(N - 1)$ fluid particles, as well as a contribution from a potential $U(\mathbf{r})$, which represents a fixed solute whose presence is felt by all N fluid particles. Since it is often convenient to distinguish between these two contributions, we will use the notation \mathbf{f}_i to denote the force on particle i arising specifically from interactions

¹ As often the case when using periodic boundary conditions, we assume that the interaction potential between pairs of particles has a finite range no greater than $L/2$, so that each particle interacts with at most one image of each other particle.

with the other fluid particles, and $-\nabla_i U(\mathbf{r}_i)$ to denote the force due to the solute; thus, $\mathbf{F}_i = -\nabla_i U(\mathbf{r}_i) + \mathbf{f}_i$.

For the special choice $\mathbf{u}(\mathbf{r}) = \mathbf{r}$, Eq.3 above reduces to a result derived by Henderson (Eq.9 of Ref.[5]; see also Ref.[6]), which relates the density of particles on the surface $\partial\Omega$ to the density within the volume Ω . In that case the last term in Eq.3 is very similar to the familiar Clausius virial[4]. More generally, however, Eq.3 encompasses a large family of virial-like theorems, each associated with a particular choice of \mathbf{u} . Such “hypervirial” relations have been studied by Hirschfelder [7], though they have been derived and expressed in a different way. We now show how, by choosing $\mathbf{u}(\mathbf{r})$ to be discontinuous over the surface of a sphere of radius R , we obtain useful identities for the fluid density averaged over this surface.

A. Hard spherical solute

To begin, let us consider a fluid in the presence of a fixed hard spherical solute. This simple solvation model plays an important role in the study of nonuniform fluid theories [2, 8, 9]. A quantity of particular interest is the *contact density*, ρ_c (the particle density in the immediate vicinity of the hard sphere), which is closely related to the free energetic cost of “growing” such a hard sphere solute in the fluid[2].

We will use a cartesian coordinate system whose origin $\mathbf{r} = \mathbf{0}$ is at the center of the box, and we assume that this point also coincides with the center of the hard solute. Let R_{sol} denote the radius of the solute, $R_{\text{max}} = L/2$ the radius of the largest sphere that fits inside the simulation box, and Ω the thick spherical region between these two radii. That is, Ω is defined by $R_{\text{sol}} < r < R_{\text{max}}$, and the volume of this region is $V_\Omega = (4\pi/3)(R_{\text{max}}^3 - R_{\text{sol}}^3)$. With the aim of determining the particle density at a distance R from the center of the solute (where $R_{\text{sol}} < R < R_{\text{max}}$), let us construct the vector field

$$\mathbf{u}(\mathbf{r}; R) = \begin{cases} \frac{r^3 - R_{\text{sol}}^3}{3r^2} \hat{\mathbf{r}}, & (R_{\text{sol}} \leq r \leq R) \\ \frac{r^3 - R_{\text{max}}^3}{3r^2} \hat{\mathbf{r}}, & (R < r \leq R_{\text{max}}) \end{cases} \quad (4)$$

where the notation $\mathbf{u}(\mathbf{r}; R)$ indicates the parametric dependence of \mathbf{u} on the radius of interest R . This field vanishes at the boundaries of Ω (i.e. at R_{sol} and R_{max}), therefore so does the term on the left side of Eq.3. The divergence of \mathbf{u} inside the region Ω is easily calculated to

be

$$\nabla \cdot \mathbf{u} = 1 - \frac{V_\Omega}{4\pi R^2} \delta(r - R), \quad (5)$$

where the delta-function arises from the discontinuity of \mathbf{u} at the surface $|\mathbf{r}| = R$. From this we get

$$\int_\Omega d\mathbf{r} \rho \nabla \cdot \mathbf{u} = \langle N_\Omega \rangle - V_\Omega \rho(R), \quad (6)$$

where $\langle N_\Omega \rangle = \int_\Omega d\mathbf{r} \rho(\mathbf{r})$ is the average number of particles found in the region Ω , and

$$\rho(R) = \frac{1}{4\pi R^2} \int d\mathbf{r} \rho(\mathbf{r}) \delta(r - R) \quad (7)$$

is the particle density, averaged over the spherical surface $r = R$. Combining this result with Eq.3 and rearranging terms finally gives us:

$$\rho(R) = \frac{1}{V_\Omega} \left[\langle N_\Omega \rangle + \beta \left\langle \sum_{i \in \Omega} \mathbf{u}_i \cdot \mathbf{f}_i \right\rangle \right], \quad (8)$$

where $\mathbf{u}_i = \mathbf{u}(\mathbf{r}_i; R)$. Note that we have replaced \mathbf{F}_i in Eq.3 by \mathbf{f}_i above; because the region Ω ($R_{\text{sol}} < r < R_{\text{max}}$) was defined to be entirely outside the region of space occupied by the solute ($r \leq R_{\text{sol}}$), the net force felt by any particle that is inside Ω includes only contributions from the other fluid particles. Eq.8 is an *exact* result that expresses the surface-averaged density $\rho(R)$ in terms of a bulk average: all the particles inside Ω contribute to the right side of Eq.8.

As a consistency check, we note that if our fluid is an ideal gas, then \mathbf{f}_i is identically zero, and Eq.8 becomes $\rho(R) = \langle N_\Omega \rangle / V_\Omega$. This is as expected: in the ideal gas limit the fluid density is independent of the distance away from the solute, and is therefore equal to the average number of particles in any given region of space, divided by the volume of that region. Thus for the more general case of a non-ideal fluid, the two terms inside the square brackets in Eq.8 can be interpreted as, respectively, an ideal gas contribution to the radial distribution function, and a virial-like correction due to interparticle interactions.

Eq.8 is equivalent to the statement that the quantity

$$\hat{\rho}(\Gamma; R) = \frac{1}{V_\Omega} \left(N_\Omega + \beta \sum_{i \in \Omega} \mathbf{u}_i \cdot \mathbf{f}_i \right) \quad (9)$$

is an *unbiased estimator* of the radial distribution function $\rho(R)$. Here, $\Gamma = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$ is a single N-particle configuration of our system, and the quantity on the right side above is well defined for any such configuration. Thus, $N_\Omega = N_\Omega(\Gamma)$ is the number of fluid particles

inside Ω , for the configuration Γ . By the term “unbiased estimator”, we mean simply that the average of $\hat{\rho}(\Gamma; R)$, over configurations Γ sampled from an equilibrium (canonical) ensemble, is identically equal to the desired density $\rho(R)$; that is, $\rho(R) = \langle \hat{\rho}(\Gamma; R) \rangle$. This is the content of Eq.8.

In the example that we have just presented, $\mathbf{u}(\mathbf{r}; R)$ was chosen so as to (a) cause the surface term in Eq. (3) to vanish, and (b) “pick out” the density at a particular distance R from the origin via a discontinuity at R . We now show how the same strategy can be applied when the solute is no longer perfectly hard. This will pave the way to the consideration of particle-particle distribution functions in homogeneous fluids.

B. General spherical solute

Consider a solute described by a fixed spherically symmetric potential $U(r)$. Recalling that $\mathbf{F}_i = -\nabla_i U(r_i) + \mathbf{f}_i$, Eq. (3) can be rewritten as

$$\oint_{\partial\Omega} d\mathbf{a} \cdot \mathbf{u} \rho = \int_{\Omega} d\mathbf{r} \rho \hat{\mathcal{L}} \cdot \mathbf{u} + \beta \left\langle \sum_{i \in \Omega} \mathbf{u}_i \cdot \mathbf{f}_i \right\rangle, \quad (10)$$

where

$$\hat{\mathcal{L}} \cdot \mathbf{u} \equiv e^{\beta U} \nabla \cdot (e^{-\beta U} \mathbf{u}). \quad (11)$$

Essentially, the solute contribution to \mathbf{F}_i in the second term on the right side of Eq.3, has been transferred to the first term on the right.

Let us now take Ω to be the entire region of space enclosed by a sphere of radius $R_{\max} = L/2$. By analogy with the hard sphere example, let us search for a vector field $\mathbf{u}(\mathbf{r}; R)$ that: (a) vanishes on the surface of Ω , (b) is discontinuous at $r = R$, and (c) is spherically symmetric and oriented along the radial direction, i.e. $\mathbf{u}(\mathbf{r}; R) = u(r; R) \hat{\mathbf{r}}$. We also want $u(0; R) = 0$, to avoid a singularity at the origin. Note that these properties do not *uniquely* specify $\mathbf{u}(\mathbf{r}; R)$. In the following two paragraphs we will construct two different fields satisfying these conditions (Eqs.13 and 17) and we will present the unbiased estimators associated with these fields (Eqs.16 and 19). In the Appendix, we show how this approach can be generalized.

With Eq.5 in mind, let us first search for a field $\mathbf{u}(\mathbf{r}; R)$ that satisfies

$$\hat{\mathcal{L}} \cdot \mathbf{u} = 1 - \alpha \delta(r - R), \quad (12)$$

where α is a constant whose value is determined by demanding that \mathbf{u} vanish at the origin and at the boundary of Ω ($r = R_{\max}$). The solution of Eq.12 is obtained by straightforward integration:

$$\mathbf{u} = \hat{\mathbf{r}} \frac{e^{\beta U(r)}}{r^2} \times \begin{cases} \int_0^r ds s^2 e^{-\beta U(s)}, & 0 \leq r \leq R \\ \int_{R_{\max}}^r ds s^2 e^{-\beta U(s)}. & R < r \leq R_{\max} \end{cases}. \quad (13)$$

This field clearly satisfies the desired boundary conditions, and by explicit evaluation we can confirm that

$$\hat{\mathcal{L}} \cdot \mathbf{u} = 1 - \delta(r - R) \frac{e^{\beta U(R)}}{4\pi R^2} Q_U, \quad (14)$$

where

$$Q_U = 4\pi \int_0^{R_{\max}} ds s^2 e^{-\beta U(s)}. \quad (15)$$

Combining Eqs.10 and 14 and carrying out the sort of rearrangement of terms that led to Eqs.8 and 9, we arrive at the following unbiased estimator for $\rho(R)$:

$$\hat{\rho}(\Gamma; R) = \frac{e^{-\beta U(R)}}{Q_U} \left(N_{\Omega} + \beta \sum_{i \in \Omega} \mathbf{u}_i \cdot \mathbf{f}_i \right). \quad (16)$$

This is very similar to Eq.9. In particular, the two terms inside the square brackets represent an ideal gas contribution to $\rho(R)$, and a virial-like correction. As in the example of the hard-sphere solute, $\hat{\rho}(\Gamma; R)$ takes on different values for different configurations Γ , but its average over an equilibrium ensemble of configurations is equal to $\rho(R)$.

Although Eq.16 is correct, the evaluation of the function $\mathbf{u}(\mathbf{r}; R)$ involves an integral of the form $\int^r ds s^2 e^{-\beta U(s)}$ (see Eq.13). In an actual application of this method this integral must generally be carried out numerically. We can avoid this inconvenience by replacing the integral in Eq.13 by an analytically tractable expression. For instance, if the solute is strongly repulsive within a region $r < R_1$, then we can replace the factor $e^{-\beta U(s)}$ appearing in the integrand, by the unit step function $\theta(s - R_1)$. We then get:

$$\mathbf{u}(\mathbf{r}; R) = \hat{\mathbf{r}} \frac{e^{\beta U(r)}}{3r^2} \left[(r^3 - R_1^3) \theta(r - R_1) + (R_1^3 - R_{\max}^3) \theta(r - R) \right] \quad (17)$$

$$\hat{\mathcal{L}} \cdot \mathbf{u} = e^{\beta U(r)} \left[\theta(r - R_1) - \frac{\delta(r - R)}{4\pi R^2} V_{\Omega^*} \right], \quad (18)$$

$$\hat{\rho}(\Gamma; R) = \frac{e^{-\beta U(R)}}{V_{\Omega^*}} \left[\sum_{i \in \Omega^*} e^{\beta U(r_i)} + \beta \sum_{i \in \Omega} \mathbf{u}_i \cdot \mathbf{f}_i \right], \quad (19)$$

where Ω^* is the spherical shell defined by $R_1 < r < R_{\max}$, whose volume is $V_{\Omega^*} = (4\pi/3)(R_{\max}^3 - R_1^3)$. This field satisfies the desired conditions listed two paragraphs ago, but the evaluation of $\mathbf{u}(\mathbf{r}; R)$ in this case no longer requires numerical integration.

Note that the replacement of $e^{-\beta U(s)}$ (in Eq.13) by $\theta(s-R)$ (to get Eq.17) simply amounts to a different choice of $\mathbf{u}(\mathbf{r}; R)$. The resulting estimator, Eq.19, is as exactly unbiased as the one given by Eq.16. In the Appendix we show that, in fact, both these estimators are special cases of a more general expression for $\hat{\rho}(\Gamma; R)$ (Eq.24).

C. Pair correlation function

We now consider how our method can be applied to the estimation of the pair correlation function $g(r)$ of a uniform fluid. As above, we imagine N identical, classical particles inside a cubic simulation box of volume $V = L^3$ and periodic boundary conditions, only now we assume that there are no solutes present. The total potential energy of a given configuration of particles is given by $\sum_{i<j} v(r_{ij})$, where $v(r)$ is the pairwise interaction potential (e.g. Lennard-Jones), r_{ij} is the distance between particles i and j , and the sum is over all pairs of particles.

For purpose of presentation, let us choose one of our N particles to act as a reference particle, and imagine for the moment that we hold the position of that particle fixed at the center of the simulation box. The pair correlation function is given by $g(r) = \rho(r)/\rho_b$, where $\rho(r)$ is the density of particles at a distance r from the reference particle, and $\rho_b = (N-1)/V$ is the bulk density of the unconstrained particles (see e.g. Ref. [3], pp. 196-197.) We can think of the reference particle as a fixed, spherically symmetric solute, which happens to be physically identical to the remaining fluid particles. Since Eqs.16 and 19 both provide unbiased estimators for the density profile around such a solute, we can apply either of those results to the present case, replacing $U(r)$ by $v(r)$ to get an unbiased estimator for $g(R)$. For instance, from Eq.19 we get

$$\hat{g}(\Gamma; R) = \frac{e^{-\beta v(R)}}{\rho_b V_{\Omega^*}} \left[\sum'_{i \in \Omega^*} e^{\beta v(r_i)} + \beta \sum'_{i \in \Omega} \mathbf{u}_i \cdot \mathbf{f}'_i \right]. \quad (20)$$

Here the notation \sum' indicates that the reference particle is not included in the sum (since that particle is being treated as a fixed solute), and similarly \mathbf{f}'_i is the net force acting on

particle i from all other particles except the reference particle. The \mathbf{u} -field that enters Eq.20 is given by Eq.17, with $U(r)$ replaced by $v(r)$.

The discussion of the previous paragraph was framed around the idea that a single reference particle is held fixed, and $g(R)$ is computed with respect to that particle alone. However, in an actual implementation of the method, all N particles would be treated equally during the simulation. For every configuration $\Gamma = (\mathbf{r}_1, \dots, \mathbf{r}_N)$ generated during such a simulation, the right side of Eq.20 would be evaluated N times, with each particle taking its turn as the reference particle.² Each of these N evaluations represents a single unbiased estimate of $g(R)$; by averaging over all N of them, we obtain better statistics than by the use of only a single reference particle. Effectively, the same procedure is used when applying the usual histogram method: instead of specifying one reference particle and compiling statistics on the relative locations of the other $N - 1$ particles, the pair correlation function g is estimated by keeping track of all $N(N - 1)/2$ pairwise distances in the fluid.

D. Generalization to NPT ensemble

So far in our presentation, we have taken the volume V of the simulation box to be constant, as appropriate for quantities computed in the NVT equilibrium ensemble. However, our method easily extends to the NPT ensemble, which can be viewed as a superposition of NVT ensembles. Explicitly, in the NPT case we have

$$\rho(R) = \langle \langle \hat{\rho}(\Gamma; R, V) \rangle \rangle = \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M \hat{\rho}(\Gamma_m; R, V_m). \quad (21)$$

The double angle brackets signify an average over both a distribution of system volumes V , and for each volume a canonical average over microstates Γ sampled at that fixed volume. The quantity $\hat{\rho}(\Gamma; R, V)$ is any one of the unbiased estimators derived above, with the dependence on V (or equivalently, L) made notationally explicit. The final term in the equation above refers to an average over M configurations generated during an NPT simulation, in which both the microstate Γ and the volume V fluctuate with time.

² The vector \mathbf{r}_i then refers to the displacement of the i 'th particle relative to the current reference particle, i.e. we temporarily treat the location of the reference particle as the origin of the coordinate system.

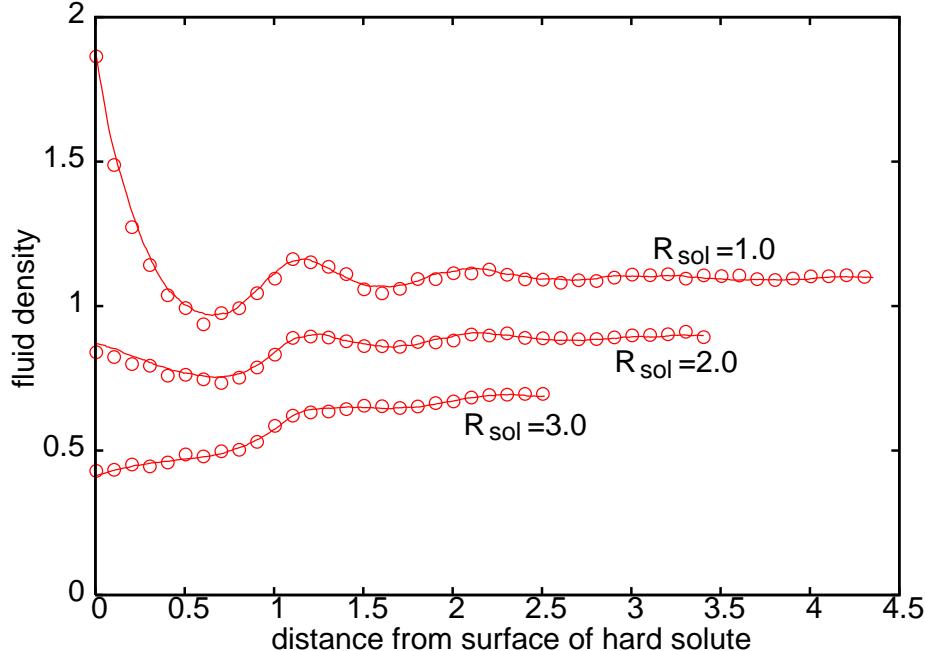


FIG. 1: Estimates of the fluid density obtained from the histogram method (circles) and with the unbiased estimator of Eq.9 (lines). For the histogram method, we used a bin size $\Delta r = 0.01$, although only every tenth bin is shown in the figure, to avoid cluttering. When using our unbiased estimator, $\rho(r)$ was evaluated at increments of 0.01, resulting in the smooth curves shown above. Finally, the curves corresponding to $R_{\text{sol}} = 2.0$ and $R_{\text{sol}} = 1.0$ have been shifted upward by 0.2 and 0.4 units, respectively, again to avoid cluttering.

III. NUMERICAL RESULTS

To test our method, we have performed Monte Carlo simulations of a truncated and shifted Lennard-Jones (LJ) fluid both in the NVT and NPT ensembles [1]. We have checked the reliability of our simulations with the available literature data on the LJ equation of state [10], always finding good agreement within statistical error bars estimated through the block average method of Flyvbjerg and Petersen [1, 11].

As a proof-of-principle test case, we first ran three simulations of a Lennard-Jones fluid surrounding a fixed hard spherical solute, in the NPT ensemble, with number of particles $N = 864$, temperature $T = 0.85$, and pressure $p = 0.022$. We used a Lennard-Jones cut-off radius $r_c = 2.5$ for the fluid particles, and the solute radius was taken as $R_{\text{sol}} = 1.0, 2.0$, and 3.0 for the three simulations. (Here and below we use the reduced LJ units [1] when

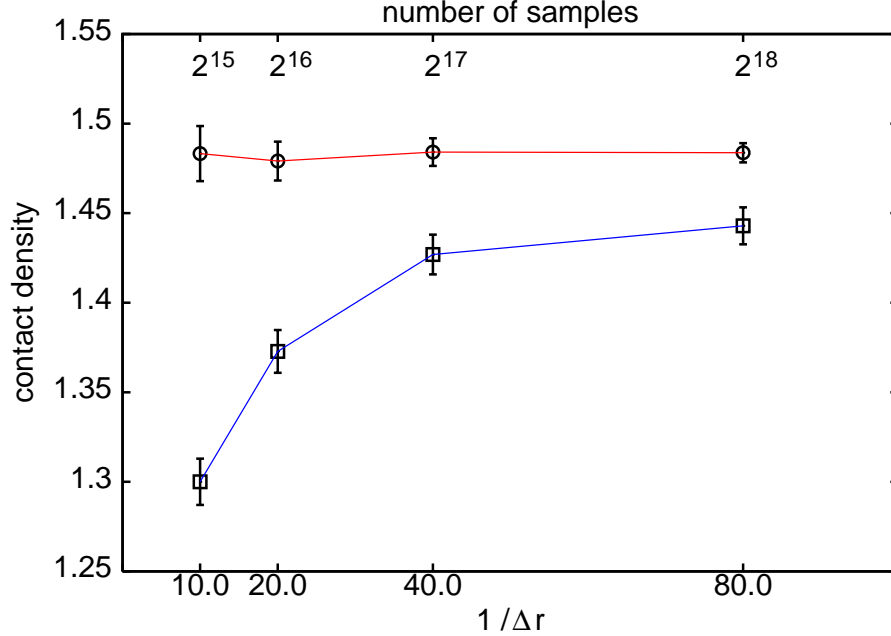


FIG. 2: Estimates of contact density at the surface of a hard cavity of radius $R_{\text{sol}} = 1.0$, obtained from four separate simulations. The lower points (squares) were obtained with the histogram method. The upper points (circles) were obtained using Eq.9. (The connecting lines are guides to the eyes.) Each simulation is identified by the number of samples generated, and by the inverse bin size used in the estimates based on the histogram method.

specifying these parameters.) These values are identical to those used to produce Fig.3 of Huang and Chandler[8]. From the data generated during each simulation (2^{13} statistically independent configurations), we estimated the radial distribution function $\rho(r)$ using the histogram method, as well as with the unbiased estimator given by Eq.9 above (with \mathbf{u} given by Eq.4). Fig.1 shows good agreement between the two.

Next, to highlight the bias due to finite bin size when using the histogram method – and to illustrate that our method does not suffer from this bias – we ran four new simulations of a solvated sphere of radius $R_{\text{sol}} = 1.0$, with the NPT parameters described in the previous paragraph. The four simulations differed in duration, generating $n_s = 2^{15}, 2^{16}, 2^{17}$, and 2^{18} samples, respectively. From each simulation we estimated the contact density $\rho_c = \rho(R_{\text{sol}}^+)$ using both the histogram method (i.e. by counting the number of particles within a distance Δr from the surface of the sphere) and the unbiased estimator of Eq.9. With the histogram method, we varied the bin size Δr from one simulation to the next, keeping the product

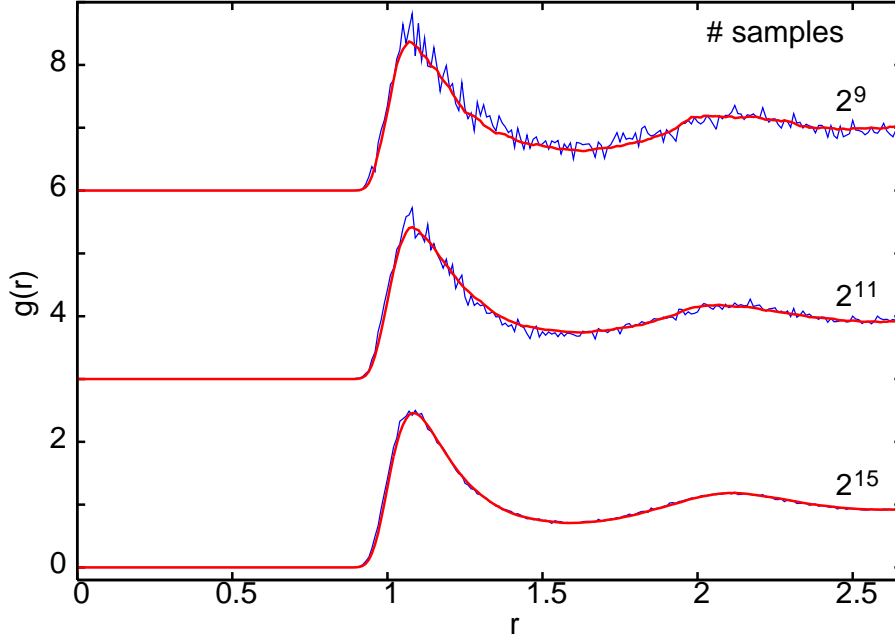


FIG. 3: The particle-particle distribution function $g(r)$ for a uniform Lennard-Jones fluid, comparing the present method (thick red lines) with the usual histogram (thin, noisy blue lines) for different numbers of samples. Each curve was evaluated at points spaced by $\Delta r = 0.01$, corresponding to the bin size used in the histogram method. The curves have been shifted vertically in order to facilitate visualization. The error bars of the present method (not shown) are always smaller than the histogram ones.

$n_s \Delta r$ constant so that the number of counts in the bin was roughly the same for each simulation. Because $\rho(r)$ increases sharply as one approaches the solute surface (see the upper curve in Fig.1), we expect the histogram method to underestimate ρ_c for any finite bin size, but this systematic error should diminish as we decrease the size of the bin. Fig.2 shows the estimates of ρ_c that we obtained from our simulations. The results obtained with the histogram method (squares) are consistent with our expectations: with decreasing bin size, the estimate of ρ_c increases, suggesting an asymptotic approach to the correct value as $\Delta r \rightarrow 0$. However, this comes at a price: more samples are needed to maintain the same level of statistical error. When using Eq.9 (circles), there is no discernable bias; the statistical error bars simply become smaller as the number of samples increases.

Finally, Fig.3 shows estimates of the pair correlation function $g(r)$ obtained from the simulation of a uniform Lennard-Jones fluid (no solute) in the NVT ensemble, with $N = 108$,

$V = N/\rho_b$ (where the bulk density was set at $\rho_b = 0.70$), and $T = 0.85$. Estimates of $g(r)$ were obtained after the generation of 2^9 , 2^{11} , and finally 2^{15} samples, using both Eq.20 and the histogram method. Although exactly the same data were used for each overlying pair of curves, our unbiased estimator furnishes a smoother estimate of $g(r)$, with the difference between the two methods diminishing with increasing number of samples.

IV. CONCLUSIONS

We have introduced and illustrated a general method for the numerical estimation of spatial distribution functions in classical fluids. Our method relies on the construction of unbiased estimators (see Eqs.9, 16, 19, 20, 24), derived from a virial-like identity (Eq.3). We have focused in this paper on cases for which the spatial distribution function of interest has spherical symmetry, reflecting e.g. the symmetry of the fixed solute, but the extension of our method to cases of oddly shaped solutes, or nonuniform averages over possibly non-spherical surfaces $\partial\Omega$, is not as difficult as might first appear to be the case³, and we hope to study this issue in greater detail in future work. Another natural extension would be in the direction of complex fluids.

While the use of simple histograms does not normally pose a serious computational challenge to the estimation of spatial distribution functions (at least for simple fluids), the numerical results of Section III suggest that our method has certain advantages, such as the elimination of systematic bias and the generation of smooth results.

Acknowledgments

The authors would like to thank Hank Ashbaugh, Gerhard Hummer, Lawrence Pratt, John Weeks, and Tom Woolf for fruitful discussions and suggestions. This research was supported by the Department of Energy, under contract W-7405-ENG-36.

³ We would like to thank Hank Ashbaugh for valuable suggestions with regard to this point.

Appendix

Here we briefly outline a generalization of the approach that led to Eqs.16 and 19. The physical context is that of a fluid surrounding a spherically symmetric solute, but as illustrated in the main text any results obtained in this case can just as well be applied to the estimation of the pair correlation function $g(r)$.

Recall that we want to find a \mathbf{u} field that: vanishes at $r = R_{\max}$; is discontinuous at $r = R$; and has the form $\mathbf{u}(\mathbf{r}; R) = u(r; R)\hat{\mathbf{r}}$, with $u(0; R) = 0$. These conditions are satisfied quite generally by the following field:

$$\mathbf{u}(\mathbf{r}; R) = \hat{\mathbf{r}} \frac{e^{\beta U(r)}}{r^2} \int_{a(r)}^r ds s^2 e^{-\beta \phi(s)} \quad , \quad a(r) = R_{\max} \cdot \theta(r - R), \quad (22)$$

where $\phi(r)$ is an arbitrary function, and $\theta(\cdot)$ is the unit step function. Because of the way that $a(r)$ is defined, $\mathbf{u}(\mathbf{r}; R)$ points *outward* (i.e. away from the origin) for $r < R$, and *inward* for $r > R$, as was the case with the fields defined by Eqs.4, 13 and 17. From Eq.22 we get

$$\hat{\mathcal{L}} \cdot \mathbf{u} = e^{\beta U(r)} \left[e^{-\beta \phi(r)} - \frac{\delta(r - R)}{4\pi R^2} Q_\phi \right], \quad (23)$$

where $Q_\phi = 4\pi \int_0^{R_{\max}} ds s^2 e^{-\beta \phi(s)}$, and $\hat{\mathcal{L}}$ is defined by Eq.11. Combining Eqs.10 and 23 and once again rearranging terms in the manner that led to Eqs.8 and 9, we get the following unbiased estimator for $\rho(R)$:

$$\hat{\rho}(\Gamma; R) = \frac{e^{-\beta U(R)}}{Q_\phi} \left[\sum_{i \in \Omega} e^{\beta(U - \phi)_i} + \beta \sum_{i \in \Omega} \mathbf{u}_i \cdot \mathbf{f}_i \right], \quad (24)$$

where $(U - \phi)_i \equiv U(r_i) - \phi(r_i)$.

The function $\phi(r)$ plays the role of a free parameter, to be chosen so as to maximize computational convenience. If we choose $\phi(r) = U(r)$, then we arrive at the estimator defined by Eqs.13 and 16. Alternatively, we can take $\phi(r)$ to be a hard spherical potential ($\phi = +\infty$ for $r \leq R_1$ and $\phi = 0$ for $r > R_1$) in which case we are led to the results given by Eqs.17 and 19.

[1] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic Press, San Diego, 2002).

- [2] J. D. Weeks, *Annu. Rev. Phys. Chem.* **53**, 533 (2002).
- [3] D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford Univ. Press, New York, 1987).
- [4] R. Becker, *Theory of Heat* (Springer-Verlag, New York, 1967), 2nd ed.
- [5] J. R. Henderson, *J. Chem. Phys.* **84**, 3385 (1986).
- [6] J. G. Powles, G. Rickayzen, and M. L. Williams, *J. Chem. Phys.* **83**, 293 (1985).
- [7] J. O. Hirschfelder, *J. Chem. Phys.* **33**, 1462 (1960).
- [8] D. M. Huang and D. Chandler, *Phys. Rev. E* **61**, 1501 (2000).
- [9] K. Katsov and J. D. Weeks, *Phys. Rev. Lett.* **86**, 440 (2001).
- [10] J. Johnson, J. Zollweg, and K. Gubbins, *Mol. Phys.* **78**, 591 (1993).
- [11] H. Flyvbjerg and H. G. Petersen, *J. Chem. Phys.* **91**, 461 (1989).